Введение в искусственный интеллект. Машинное обучение

Лекция 4. Линейная регрессия

Бабин Д.Н., Иванов И.Е., Петюшко А.А.

кафедра Математической Теории Интеллектуальных Систем

10 марта 2020г.

Линейная регрессия

Постановка задачи и допущения

- $X = \mathbb{R}^n$, $Y = \mathbb{R}$
- $a(x) = f_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$, где $\theta = (\theta_0, \theta_1, \ldots, \theta_n)^T \in \mathbb{R}^{n+1}$ — параметры модели.
- Удобно писать в векторном виде

$$a(x) = \theta^T \cdot x,$$

где $x = (x_0, x_1, \ldots, x_n)^T$ и $x_0 = 1$.

Метод наименьших квадратов

- $L(\theta, X_{train}) = MSE(\theta, X_{train}) = \sum_i (\theta^T \cdot x^{(i)} - y_i)^2$ — функция потерь
- Задача найти $\hat{\theta} = \arg\min_\theta (L(\theta, X_{train}))$

Аналитическое решение

Теорема

Решением задачи $\arg\min\limits_{\theta}(\sum\limits_{i=1}^{\ell}(\theta^T \cdot x^{(i)} - y_i)^2)$ является $\hat{\theta} = (X^T X)^{-1} \cdot X^T \cdot y$, где $X_{i,j} = x_j^{(i)}$, $y = (y_1, ..., y_\ell)$.

Доказательство

Запишем задачу в векторном виде $||X\theta - y||^2 \to \min\limits_{\theta}$. Необходимое условие минимума в матричном виде имеет вид:

$$\frac{\partial}{\partial\theta}||X\theta - y||^2 = \frac{\partial}{\partial\theta}\left((X\theta - y)^T \cdot (X\theta - y)\right) = 2X^T(X\theta - y) = 0,$$

откуда получаем $X^T X\theta = X^T y$, из чего и следует требуемое.

Леммы

Определение

Пусть $\theta = (\theta_1, ..., \theta_n)$ — вектор столбец, а $z = z(\theta_1, ..., \theta_n)$. Тогда определим

$$\frac{\partial z}{\partial \theta} := (\frac{\partial z}{\partial \theta_1}, ..., \frac{\partial z}{\partial \theta_n})^T$$

Лемма 1

$$\frac{\partial}{\partial x} x^T a = a$$

Лемма 2

$$\frac{\partial}{\partial x} x^T A x = (A + A^T) x$$

Вероятностная интерпретация

Модель шума

$$y(x_i) = f_\theta(x_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Метод максимума правдоподобия

$$L(\varepsilon_1, ..., \varepsilon_n|\theta) = \prod_i \frac{1}{\sigma_i\sqrt{2\pi}} exp(-\frac{\varepsilon_i^2}{2\sigma_i^2}) \to \max_\theta$$

$$-L(\varepsilon_1, ..., \varepsilon_n|\theta) = const(\theta) + \frac{1}{2}\sum_i \frac{1}{\sigma_i^2}(f_\theta(x_i) - y_i)^2 \to \min_\theta$$

Преимущества и недостатки линейной регрессии

Преимущества

- Простой алгоритм, вычислительно не сложный
- Линейная регрессия хорошо интерпретируемая модель
- Несмотря на свою простоту может описывать довольно сложные зависимости (например, полиномиальные)

Недостатки

- Алгоритм предполагает, что все признаки числовые
- Алгоритм предполагает, что данные распределены нормально, что не всегда так
- Алгоритм сильно чувствителен к выбросам

Гребневая регрессия

L2-регуляризация

- $L(\theta, X_{train}) = MSE(\theta, X_{train}) + \frac{\alpha}{2}\sum\limits_{i=0}^{n}\theta_i^2 = \sum\limits_{i}(\theta^T\cdot x^{(i)} - y_i)^2 + \frac{\alpha}{2}\sum\limits_{i=0}^{n}\theta_i^2$ — функция потерь
- Задача найти $\hat{\theta} = \arg\min\limits_{\theta}(L(\theta, X_{train}))$

Теорема

Решением задачи $\arg\min\limits_{\theta}(\sum\limits_{i=1}^{\ell}(\theta^T\cdot x^{(i)} - y_i)^2 + \frac{\alpha}{2}\sum\limits_{i=0}^{n}\theta_i^2)$ является $\hat{\theta} = (X^TX + \alpha I_n)^{-1}\cdot X^T\cdot y$, где $X_{i,j} = x_j^{(i)}$, $y = (y_1, \ldots, y_\ell)$, I_n — единичная матрица.

Доказательство

Аналогично

LASSO

L1-регуляризация

- $L(\theta, X_{train}) = MSE(\theta, X_{train}) + \alpha \sum_{i=0}^{n} |\theta_i| = \sum_{i} (\theta^T \cdot x^{(i)} - y_i)^2 + \alpha \sum_{i=0}^{n} |\theta_i|$ — функция потерь
- Задача найти $\hat{\theta} = \arg\min_{\theta}(L(\theta, X_{train}))$

Свойства

- Эта регуляризация обеспечивает отбор признаков

Elastic Net

L1-регуляризация и L2-регуляризация

- $L(\theta, X_{train}) = MSE(\theta, X_{train}) + r\alpha \sum_{i=0}^{n} |\theta_i| + (1-r)\frac{\alpha}{2} \sum_{i=0}^{n} \theta_i^2 =$
  $\sum_i (\theta^T \cdot x^{(i)} - y_i)^2 + r\alpha \sum_{i=0}^{n} |\theta_i| + (1-r)\frac{\alpha}{2} \sum_{i=0}^{n} \theta_i^2$ — функция потерь
- Задача найти $\hat{\theta} = \arg\min_{\theta}(L(\theta, X_{train}))$

Свойства

- Нет аналитического решения

Вероятностный смысл регуляризации

Дано

- Параметрическая модель плотности распределения $p(x, y|w)$
- Априорная информация о плотности распределения параметров модели $p(w)$
  Например, параметрическое семейство априорных распределений $p(w; h)$, где h — неизвестная и неслучайная величина (гиперпараметр).

Тогда:

- Плотность $p(X^m, w; h) = p(X^m|w)p(w; h)$
- Тогда правдоподобие будет иметь вид

$$L(w, X^m) = \ln p(X^m, w; h) = \sum_{i=1}^{m} \ln p(x_i, y_i|w) + \ln p(w; h) \to \max_{w,h}$$

Вероятностный смысл регуляризации

$$L(w, X^m) = \ln p(X^m, w; h) = \sum_{i=1}^{m} \ln p(x_i, y_i | w) + \ln p(w; h) \to \max_{w,h}$$

- Первое слагаемое – логарифм правдоподобия (зависит только от данных);
- Второе слагаемое – логарифм априорного распределения параметров модели (зависит только от гиперпараметра);
- Второе слагаемое имеет смысл аддитивного **регуляризатора**.

Рассмотрим подробнее простые методы регуляризации.

Вероятностный смысл L1 и L2 регуляризации

Пусть веса w_i - независимые с.в. с нулевым средним и дисперсией $D > 0$.

L_2-регуляризация

Пусть w распределены нормально: $p(w; D) = \frac{1}{(2\pi D)^{\frac{n}{2}}} \exp\left(-\frac{||w||^2}{2D}\right)$.
Логарифмируем: $-\ln p(w; D) = \frac{1}{2D}||w||^2 + const(w)$

L_1-регуляризация

Пусть w распределены согласно распределению Лапласа: $p(w; D) = \frac{1}{(2D)^n} \exp\left(-\frac{||w||}{D}\right)$.
Логарифмируем: $-\ln p(w; D) = \frac{1}{D}||w|| + const(w)$

Пусть $\tau = \frac{1}{D}$ – коэффициент регуляризации.
Вероятностный смысл параметра регуляризации: τ обратно пропорционален дисперсии D вектора параметров. Увеличивая параметр τ, уменьшаем дисперсию вектора параметров $\Rightarrow$ запрещаем коэффициентам w_i принимать слишком большие значения.

Заключение

- Линейная регрессия — простая, хорошо интерпретируемая модель, не устойчивая к выбросам
- Имеет наглядную вероятностную интерпретацию
- Регуляризация — отличный способ борьбы с переобучением и шумом в данных